Deepfake Detection Using Deep Learning

Hu, Michelle (School: Woodside High School)

Deepfakes are a new machine learning technology that allow users to place the face of one person in an image onto the head of a person in another. They pose a large cybersecurity risk because they allow bad actors to spread disinformation for the purposes of election manipulation, scamming, blackmail, and more. The purpose of this project was to use deep learning, a subset of machine learning that involves teaching complex models to identify and classify types of data, to combat this growing threat. To do this, I created several deepfakes and a classification-based deep learning model trained on the FaceForensics dataset, a dataset for deepfake detection, then tested how accurately the model could classify the deepfakes when mixed with real and fake images from the FaceForensics dataset. I found that my deep learning model had very low prediction accuracies for my deepfakes but high prediction accuracies for the FaceForensics images. This was likely because the model was trained on FaceForensics images, which were not representative of actual deepfakes such as mine. In conclusion, I achieved my goal of using deep learning to detect deepfakes, and my results showed that both deepfake detection tools, such as deep learning, and resources, such as the FaceForensics dataset, require more improvement until they can be reliably used. Recommendations for future research include changing my datasets or testing different deep learning models.

Awards Won: